大数据环境中交互式查询差分隐私保护模型 *

王 迪，袁 健，申泽宇

(上海理工大学 光电信息与计算机工程学院，上海 200093)

摘 要：随着大数据时代的到来，数据挖掘技术被广泛应用，而线性查询作为该技术中最基础和最频繁的操作，其隐私保护在数据分析和数据发布隐私保护中占有极其重要的位置。交互式线性查询的交互增加了数据的处理量，运用传统的隐私保护模型效率较低。为了解决大数据环境中交互式查询差分隐私保护问题，模型针对大规模数据集中交互式线性查询差分隐私保护的特点，通过数据关联性分析减少冗余信息，采用交替方向乘子法对查询负载矩阵进行分解，利用自适应加噪技术产生差分隐私保护所需要的合理数量的噪声，设计并行处理方法实现该模型的计算。实验将提出的模型与以往模型进行对比。结果表明，所提出的模型在提升隐私保护精度的同时，也极大地提高了算法性能，因此模型切实可行。

关键词：线性查询；差分隐私；矩阵机制；关联性分析；交替方向乘子法

中图分类号：TP309.2　　doi: 10.3969/j.issn.1001-3695.2017.12.0822

Interactive queries differential privacy protection model in big data environment

Wang Di, Yuan Jian, Shen Zeyu

(*School of Optical Electrical& Computer Engineering*, *University of Shanghai for Science& Technology*, *Shanghai* 200093, *China*)

Abstract: With the arrival of the era of big data, data mining technology is widely used, and the most basic and frequent operation of the technology, linear query, whose privacy protection occupies an extremely important position in data analysis and data release privacy protection. The data processed become more when querying in an interactive linear queries way, and it is less efficient when using the traditional privacy protection models. In order to solve the problem of differential privacy protection for interactive queries in big data environment, the model reduced the redundant information through data correlation analysis, decomposed the query load matrix by adopting alternating direction multiplier method, produced a reasonable amount of noise required for differential privacy protection using the adaptive noise injection technology, and a parallel processing method designed calculated it against the characteristics of interactive linear query differential privacy protection for large-scale data set. Experiment compared the model proposed to previous works. The result showed that the model proposed promoted the accuracy of privacy protection and algorithm performance greatly. Therefore, the model is feasible.

Key words: linear query; differential privacy; Matrix mechanism; frequent pattern mining; alternating direction multiplier method

0 引言

随着互联网的不断普及，大量数据可借助于采集设备和计算机网络来收集，人类进入了"数据化"的时代。大数据中蕴涵着巨大的价值，因此对大数据的挖掘和分析成为了许多单位和研究机构改进产品功能，提升服务质量，进行科学研究的有效手段。但是数据在被研究的过程中，可能在很多环节被泄露出去，造成民众隐私的泄露，甚至有可能成为电信、信用卡诈骗的信息来源，因此对信息的隐私保护是一个重要的亟待解决的课题。

隐私保护研究所针对的领域很多，数据查询隐私保护是其中的一种。而线性查询是数据查询中最基础最频繁的操作。线性查询主要有两种方式：非交互式和交互式。对于有保密性要求的线性查询，一般采用交互式线性查询方式。交互式线性查询的交互增加了数据的处理量，因此大规模数据集中的交互式线性查询的规模极大。在对其进行隐私保护时，若仍然采用以往的线性查询隐私保护方法，数据处理时间长，效率低。因此针对大规模数据集的交互式线性查询的特点，研究高效的隐私

收稿日期：2017-12-12；**修回日期：**2018-02-08　　**基金项目：**国家自然科学基金资助项目（61775139）

作者简介：王迪（1993-），女，河南商丘人，硕士，主要研究方向为隐私保护、数据挖掘（wendy_wangdi545@163.com）；袁健（1971-），女，副教授，博士，主要研究方向为数据挖掘、网络安全、图像处理、智能交通、隐私保护等；申泽宇（1990-），男，硕士，主要研究方向为隐私保护、数据挖掘.

保护模型是极其重要且具有实用价值的工作。

本文针对大规模数据集中线性查询交互式的特点，提出了大数据环境中交互式查询差分隐私保护模型。该模型通过数据关联性分析来减少由于大规模线性查询而产生的冗余信息；采用交替方向乘子法对查询负载矩阵进行分解，提高查询速度；结合用户权限，进行自适应加噪，解决数据高灵敏度带来的问题；最终，设计了并行处理算法快速完成计算，提高算法性能。

实验在单机为Inter(R) CoreTM i7 CPU 3.4 GHz，8 GB 内存，Window7 操作系统的环境下进行，采用 Webdocs 数据集，并从中选取 160，000 条数据作为原始数据集，选用开源的 Hadoop 来实现大数据环境中交互式查询差分隐私保护模型(interactive queries differential privacy protection model in big data environment，IQDPPBD)。另外，以算法的运行时间作为指标来衡量模型的算法性能，采用分布式集群验证模型的并行的可行性和扩展性，从数据可用性和算法性能角度将 IQDPPBD 与已有模型作对比。实验结果表明，算法性能与传统隐私保护算法相比效果更佳。

1　相关工作

在 20 世纪 60 年代末，统计学家 Dalenius 首次提出了隐私保护的问题。隐私保护是指任何用户（包括合法用户和潜在的攻击者）在访问数据库时，都无法获取到任意一个用户的确切信息[1]。目前的研究中主要有两种算法，一种是分组隐私保护算法，另一种是差分隐私[2,3,4,5]保护算法。其中，在隐私保护领域影响很深并且已经被广泛研究的 k-匿名[6,7,8,9,10]及其改进算法即是分组隐私保护算法，其主要思想是将数据集进行敏感数据标志，将这些标志的数据进行泛化和压缩处理，从而达到隐藏的目的。分组隐私保护算法主要存在两个问题：在一致性和背景知识攻击下，会造成敏感属性泄露[11]；无法提出一种有效且严格的方法来证明其隐私保护水平。鉴于此，差分隐私保护算法被提出。该算法能保证在攻击者拥有最大背景知识的条件下，用户仍可以抵抗各种形式的攻击[12]。差分隐私保护算法被认为是目前最强的一种隐私保护算法[13]，作为一个数学概念，它定义了一个极为严格的攻击算法，并对隐私泄露风险给出了严谨、定量化的表示和证明。差分隐私保护在大大降低隐私泄露风险的同时，极大地保证了数据的可用性。目前，已有一些研究者将差分隐私应用到一些具体问题中，并取得了一些进展。一些学者研究了差分隐私在社交媒体[14,15]、私有位置保护[16]、分布式估计中保留特征之间的隐私问题[17]、频繁模式中的隐私泄露问题[18]以及隐私保护决策树建模[19]的应用，而分布式线性查询差分隐私保护目前仍是一个有待探究的课题，具有可研究性。

对数据集的线性查询广泛存在于统计分析中，其使用频率很高，主要有两种查询方式：非交互式和交互式。非交互式线性查询先发布数据集，然后响应用户的查询请求，查询属性之间无关联关系[20]。而交互式线性查询，因其具有多次交互性，通常将所有查询请求整合成批量线性查询进行隐私保护[20]。Li 等人[21] 基于 Haar 小波机制和分层机制提出了矩阵机制，研究寻找策略矩阵来对查询矩阵降维减少添加噪声的数量，从而提高处理批线性计数查询的效率，然而其算法的复杂度为 $O(m^3N^3)$ (m 为查询的数量， N 为数据的维度)，而且其收敛的迭代次数不能确定，其近似算法的近似因子高达 $O(\sqrt{N})$ ，随后其优化算法[22]将 ε -差分隐私放宽到 (ε,δ) -差分隐私来提高效率，但是其效果仍然无法令人满意。Yuan 等人[23]提出了低秩机制，通过分解负载矩阵 $\boldsymbol{W}_{m\times n}=\boldsymbol{B}_{m\times r}\boldsymbol{L}_{r\times n}$ 来优化矩阵机制，来弥补矩阵机制的不足。与矩阵机制和拉普拉斯机制添加噪声的方式不同，该机制对中间结果 $\boldsymbol{L}$, $\boldsymbol{D}$ 添加噪声，并通过二次规划以及梯度下降对其进行求解，使其线性收敛。然而其没有考虑数据本身之间的关联性以及矩阵的求解方法不能在分布式系统中使用的情况。Hardt 等人[24]基于凸多面体的均匀采样，提出了近似最优的 K-norm 机制与 NIM 机制，但是其算法实现复杂度较高，大数据环境下均匀采样效率不高。Bhaskara 等人[25]改进了 NIM 机制理论，但其是在牺牲精度的情况下，提高算法的处理效率。

以上方法都是在查询结果上添加噪声，其误差与数据集的规模无关，只与查询的数量有关。而 Li 等人[26]利用拉普拉斯机制对同一聚类中的记录数量 num 和属性向量之和 sum 进行加噪，以此来扰动 num 和 sum，使聚类分析的结果满足差分隐私，以此来实现隐私保护。虽然该加噪方法突破了以往研究中基于结果加噪方法的桎梏，然而其添加噪声的数量却是随机的，从而致使噪声的添加量不甚合理。而本文提出的基于差分隐私的自适应加噪模型则将添加的噪声量与用户权限结合起来，提出 ε 的上界公式，实现自适应加噪，其既避免了以往基于结果加噪的弊端，又提出了科学合理的加噪方法。与已有的加噪方法相比，其具有一定的突破性和创新性。

文献[27~30]研究如何生成合成的数据集来实现线性查询。误差随着查询数量呈对数增长趋势，而且误差会随着数据集规模的增大而增大，其误差通常正比于 $O(\sqrt{n}\ln N\ln m)$ (n 为数据集的数量)。因此，此类方法只适用于小型数据集。

以往算法考虑的问题主要是如何降低隐私预算和减小误差，忽视了交互式数据处理的时效性。本文引入了关联性分析对数据进行无关处理，减少了数据量，从而减小了误差。同时采用了交替方向乘子法对矩阵分解，以便并行处理。另外，通过自适应加噪技术添加合理数量的噪声，改进了在查询结果上添加噪声的弊端，使其适用于大规模数据集。据此，本文提出了大数据环境中交互式查询差分隐私保护模型(IQDPPBD)。

2　相关定义

定义 1　线性查询。返回数据的线性组合结果的查询方式称为线性查询，如式（1）所示。

$$q_1=\chi_1+\chi_2+\chi_3+\chi_4+\cdots+\chi_i \tag{1}$$

其中 χ_i 代表要查询的属性。

定义 2　隐私保护机制。对于一个有限域 R ，从 R 中查询

到的 r 的集合组成数据集 D，$r \in R$，其样本量为 n，属性的个数为维度 d 。用算法 G 对 r 或 r 中某个维度的值进行处理，使其满足隐私保护的条件，这一过程称为隐私保护机制。

定义 3　批查询优化。具有隐私保护的批查询是指利用集合 $Q=\{q_1,q_2,\cdots\}$ 来表示一组查询，然后对每一条查询添加噪声。批查询优化指利用每条查询之间的相关性来降低需要添加的噪声，如有三个批查询：$q_1=\chi_1+\chi_2+\chi_3+\chi_4$；$q_2=\chi_1+\chi_2$；$q_3=\chi_3+\chi_4$；最优的做法是先响应 q_2 和 q_3，然后利用其结果之和来回答 q_1。

定义 4　邻近数据集(adjacent dataset)。定义数据集 D 和 D' 有完全相同的数据维度，不相同的记录的条数记为 $D\Delta D'$，$|D\Delta D'|$ 表示 $D\Delta D'$ 中记录的数量，其中 $|D\Delta D'|=1$，此时，称 D 和 D' 是邻近数据集。

定义 5　差分隐私。对于给定邻近数据集 D 和 D'，若存在随机算法 M，使得 D 和 D' 的任一输出结果在添加噪声后的概率密度 $P_r[M(D)=O]$ 满足式（2），则称算法 M 满足 ε -差分隐私。其中 O 表示可能的输出集合，ε 表示隐私预算。

$$P_r[M(D)=O] \le \exp(\varepsilon)\times P_r[M(D')=O] \tag{2}$$

从式（2）可以看出，隐私预算 ε 越小，隐私保护的程度越高。

定义 6　全局敏感度。定义邻近数据集 D 和 D'，对于任意的查询函数 $f：D\rightarrow Rd$，其全局敏感度 Δf 如式（3）所示。

$$\Delta f=\max_{D,D'}\|f(D)-f(D')\|_p \tag{3}$$

其中：R 表示函数所映射的实数空间，d 表示函数 f 的查询维度，p 表示度量 Δf 使用的 L_p 距离。

定义 7　拉普拉斯机制[24]。通过拉普拉斯分布产生的噪声来对产生的输出值进行扰动，其差分隐私保护函数如式（4）所示。

$$M(D)=F(D)+lap(\Delta f/\varepsilon) \tag{4}$$

其中：$lap(\Delta f/\varepsilon)$ 表示添加的拉普拉斯噪声。

3　大数据环境中交互式查询差分隐私保护模型 IQDPPBD

交互式查询系统中，用户的查询数据具有相关性，故而存在大量的数据冗余。关联性分析去除用户查询数据中的关联数据可以降低冗余数据。该方法在大规模数据处理中效率显著。另外，在交互式查询中，用户希望系统能快速响应查询请求，故需设计良好的算法来提高交互式系统的响应速度。为此，需要对查询数据进行并行处理，从而节省数据的处理时间，提高交互式系统的响应速度。然而，隐私保护和数据可用性往往不可兼得，针对隐私保护，添加拉普拉斯噪声实现差分隐私保护；针对数据可用性，首先结合 ε 的上界和用户的特点选择差分隐私保护中恰当的 ε 值，然后通过减少添加噪声的数量来提高数据可用性。综合考虑以上的一系列问题，本文提出了大数据环境中交互式查询差分隐私保护模型。

3.1　大数据环境中交互式查询差分隐私保护模型概述

大数据环境中交互式查询差分隐私保护模型（IQDPPBD）的核心思想是首先从数据的角度筛选出关联属性，从而对负载矩阵进行数据无关性处理，同时采用并行算法提高交互式系统的响应速度，其次结合交替方向乘子法实现分布式负载矩阵分解，最后自适应选取 ε 值实现拉普拉斯加噪，再将原去除的数据属性还原，返回完整的查询结果。

IQDPPBD 模型如图 1 所示，主要包括三个子模型：

a) 基于关联性分析的数据无关性处理模型（data-independent processing model based on correlation analysis，DPMCA）；

b) 并行梯度下降矩阵分解模型（parallel gradient matrix decomposition model，PGMDM）；

c) 基于差分隐私的自适应加噪模型（adaptive noise model based on differential privacy，ANMDP）。

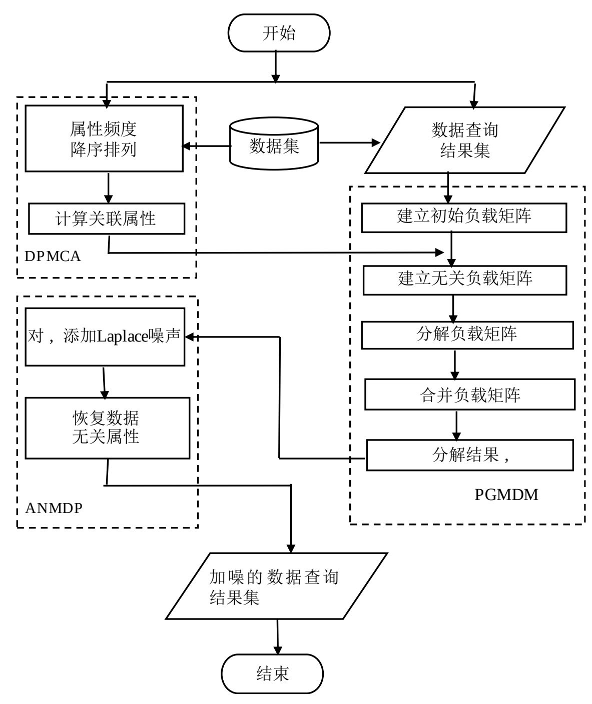


图 1　大数据环境中交互式查询差分隐私保护模型(IQDPPBD)工作流程图

IQDPPBD 模型的工作流程如下：

a) 查询数据集。

b) 采用 DPMCA 模型，计算负载矩阵数据之间的关联性，并通过设定最小支持度，选取关联性来消除负载矩阵之间的数据相关性。

c) 采用 PGMDM 模型，通过对数据无关的负载矩阵进行分割。并行执行矩阵分解算法，将由批查询数据组成的负载矩阵进行分解。

d)采用 ANMDP 模型，对 L 和 D 添加拉普拉斯噪声，实现数据集的差分隐私保护。其中 L 表示负载矩阵分解结果，D 表示数据集。

e) 将查询后的添加了噪声的结果返回给用户。

3.2　基于关联性分析的数据无关性处理模型 DPMCA

1）FP-growth 算法概述

FP-growth 算法由 Han 等人首次提出[20]，该算法为了减少对原数据集的读取次数及候选频繁项集的个数，提高挖掘效率，以共享前缀的方式在内存中构造 FP-tree 来对原始数据集进行深度压缩。构造 FP-tree 之后，频繁项集的挖掘就可以在内存中利用 FP-tree 采用频繁项目增长的方式进行，这是减少读取次数和候选频繁项集的个数的关键技术。

2）DPMCA 模型描述

数据集中有许多隐藏的数据关联模式，利用 FP-growth 算法挖掘出这种关联模式，通过挖掘出的关联模式去除查询数据中的冗余数据。

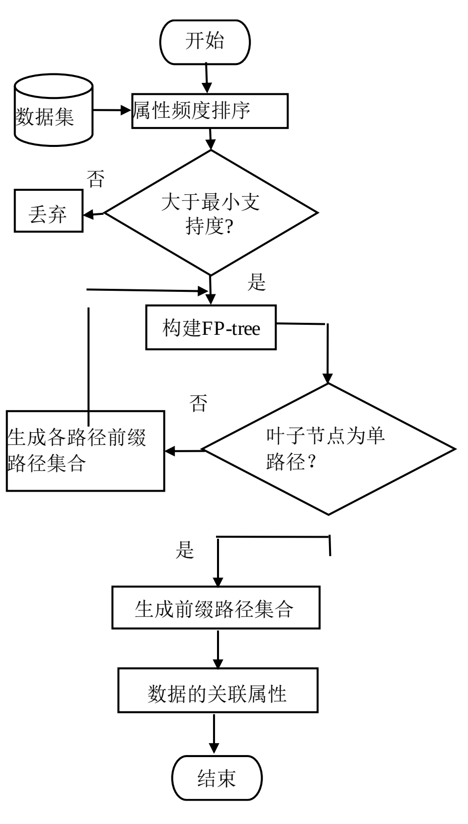


图 2　基于关联性分析的数据无关性处理模型(DPMCA)描述示意图

模型描述如下：

a) 对数据集进行扫描，得出每一个属性的频度，按照属性频度进行降序排列，得到属性频度降序列表。

b) 设定最小支持度 n ，根据数据属性频度降序列表，筛选掉频度小于最小支持度 n 的属性。

c) 构建 FP-tree[20]。将排好序的数据集插入到前缀树中，构成 FP-tree。同时对第一次出现的节点建立链表。

d) 用 FP-Growth 算法对 FP-tree 进行整理。

e) 如果叶子节点为单路径，去除叶子节点，生成前缀路径的集合，进入 Step6。如不为单路径，生成各路径前缀路径的集合，构成一个新的 FP-tree，返回 d)。

f) 获取生成的前缀路径的集合，即为查询记录的关联属性。

3.3　并行梯度下降矩阵分解模型 PGMDM

分析交互式查询系统的特点，发现其数据量大，查询回合多，因此要求查询算法收敛速度快，单词查询时间短。本文基于交替方向乘子法（alternating direction method of multipliers，ADMM）和低秩机制提出 PGMDM 模型，用来提高差分隐私矩阵分解的效率。

交互式查询系统中，用户查询为批量线性查询，属于统计学习问题之一。它首先把通过初始查询结果构建的负载矩阵剔除，然后根据 DPMCA 模型得出的数据无关属性得到无关负载矩阵，最后再进行矩阵分解。利用 Yuan[23]等人提出的低秩机制式（5）（6）分别计算分解矩阵 $\boldsymbol{B}$ 和 $\boldsymbol{G}$ 相对于 L 的梯度 $\frac{\partial \boldsymbol{G}}{\partial L}$ 。

$$\boldsymbol{B}=(\beta \boldsymbol{W}L^{T}+\pi L^{T})(\beta LL^{T}+\boldsymbol{I})^{-1} \tag{5}$$

$$\frac{\partial \boldsymbol{G}}{\partial L}=\beta \boldsymbol{B}^{T}\boldsymbol{B}L-\beta \boldsymbol{B}^{T}\boldsymbol{W}-\boldsymbol{B}^{T}\pi \tag{6}$$

其中 $\boldsymbol{B}$ 和 L 分别表示对负载矩阵 $\boldsymbol{W}$ 分解矩阵后的两个矩阵。式(5)用来更新 $\boldsymbol{B}$ ，且 $\boldsymbol{B}$ 的计算只与 L 的更新有关。

该模型结合矩阵的特性，将负载矩阵 $\boldsymbol{W}$ 分解成多个矩阵，分发到各个节点上计算。其过程如图 3 所示。

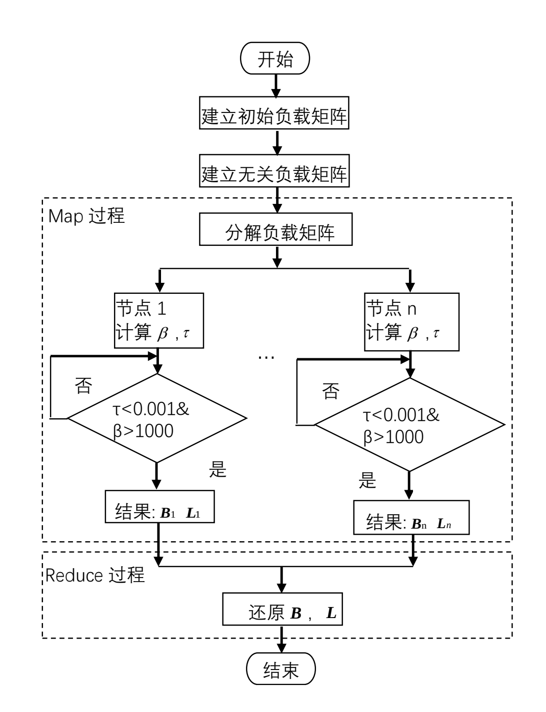


图 3　并行梯度下降矩阵分解模型(PGMDM)描述示意图

模型的描述如下：

a) 根据用户的查询要求生成初始结果负载矩阵。

b) 根据 DPMCA 模型得出的关联属性对负载矩阵进行初始化处理，将初始负载矩阵中关联数据进行转换剔除生成无关负载矩阵。

c) 对无关负载矩阵分解以便并行处理。将无关负载矩阵 $\boldsymbol{W}_{m\times r}$ 分解成 n 部分，每一部分行数为 $\frac{m}{n}$ ，列数为 r 。 n 为分布式系统的节点数。

引入分布式计算的 Map 过程：首先访问数据集，遍历每一行数据，记录行号 L，然后对输出的 key 值取整作为分组号 $\frac{L}{n}$，令 value 为数据中的 $\frac{m}{n}$ 行数据。Combiner 过程：聚合每组中的数据，形成待处理的数据。将划分过的部分，分发到 n 个节点上。

d) 在每个节点上计算分解因子，并更新 β 和 $\tau=\|W-BL\|$。当 β 大于 1000 且 τ 小于 0.001 时停止迭代。

引入云计算中的 Reduce 过程：将各节点计算出的 B_i, L_r 以及组号 $\frac{L}{n}$ 写入云计算的 Reduce 过程实现整合。将相同组号的 L_r 按行号进行拼接，得到完整的 L。

3.4　基于差分隐私的自适应加噪模型 ANMDP

由差分隐私的定义可知要满足差分隐私，必须添加符合拉普拉斯分布的噪声。然而，数据灵敏度高，直接增加噪声会导致数据不可用[31]，所以需要寻求一种合适的加噪方法。而隐私保护程度取决于差分隐私中 ε 的选取，ε 越小隐私保护程度越强但添加的噪声量越大，ε 越大隐私保护程度越弱且添加的噪声越少。因此，选取合理的 ε 将有助于兼顾隐私保护程度和噪声量。ANMDP 模型在考虑 ε 的上界 $\varepsilon \le \frac{\ln 2(1-\rho)\Delta q}{L}$ 的情况下结合用户的权限，实现自适应对 PGMDM 算法中得出的矩阵 $\boldsymbol{L}$，$\boldsymbol{D}$ 添加拉普拉斯噪声。其中，用户的权限越高 ε 的选取越接近上界，权限越低选取的 ε 越小。

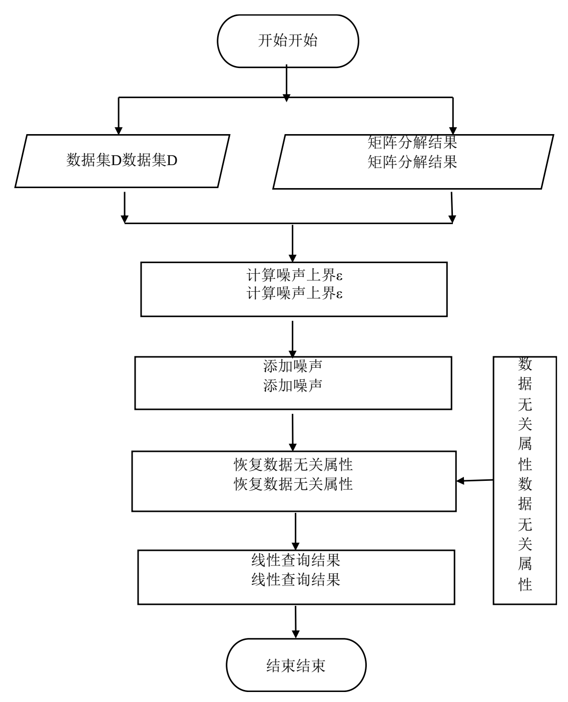


图 4　基于差分隐私的自适应加噪模型(ANMDP)描述示意图

模型的描述如下：

a) 由 $\varepsilon \le \frac{\ln 2(1-\rho)\Delta q}{L}$ 计算出 ε 的上界，结合用户的权限选择 ε。

b) 利用拉普拉斯机制[24]对 $\boldsymbol{L}$, $\boldsymbol{D}$ 添加满足 ε 的噪声。

c) 将去掉的数据无关的属性进行还原。

d) 返回给用户查询结果。

4　实验结果与分析

实验从数据可用性和算法性能角度对比 IQDPPBD、LRM[23] 和 MM[22] 模型。实验采用 Frequent Itemset Mining Data Repository[30]的 webdocs 数据集，该数据集每条记录代表顾客的购买行为，把顾客购买的一个商品作为一个数据项。选取其中 160，000 条数据作为原始数据集，随机生成 m 个批查询。为了验证并行性同时考虑可扩展性，采用分布式集群验证实验。常用的分布式集群实验平台主要有 MPI 和 Hadoop，但由于 MPI 存在节点失效，网络通信故障等问题，本实验选用开源的 Hadoop 来实现 IQDPPBD，并用算法的运算时间来衡量算法性能。本实验模型用 java 语言实现。由于差分隐私添加噪声具有一定的随机性，最终结果取 20 次实验结果的平均值。实验单机环境为 Inter(R) Core(TM) i7 CPU 3.4GHz，8GB 内存，window7 操作系统，分布式环境如表 1 所示。

表 1　软硬件配置表

Host	IP	OS	CPU	Memory	JDK
Master	192.168 1.100	CentOS 7.0	双核 1.8GHz	2G	JDK_1.8.0 _111
Slaver -1	192.168 1.101	CentOS 7.0	双核 1.8GHz	2G	JDK_1.8.0 _111
Slaver -2	192.168 1.102	CentOS 7.0	双核 1.8GHz	2G _111	JDK_1.8.0
Slaver -3	192.168 1.103	CentOS 7.0	双核 1.8GHz	2G	JDK_1.8.0 _111

4.1　实验评价指标

1）数据可用性

为了衡量噪声的干扰程度，并以此得出查询结果的精确程度，本实验引入欧氏距离。传统的隐私保护算法[22,23] 需事先指定添加的噪声量，而 IQDPPBD 模型与此不同，其关键技术在于自适应添加噪声，这就需要针对不同的 ε 值测量噪声干扰程度，从而得出查询结果的精确度。本实验将 ε 分别设置为 1.25，1.0，0.75 和 0.5，其中，ε 的前两项对应较高权限用户，后两项对应较低权限用户。另外，查询规模 m 设置为 5000 条。

需要说明的是，欧氏距离（L_2 范数）是两点或多点之间的距离表达式，如式（7）所示，距离值 d 越小说明与原数据差别越小，噪声干扰越小，查询精确度越高。

$$\mathrm{d}=\sqrt{\sum_{k=1}^{n}(\chi_{1k}-\chi_{2k})^2} \tag{7}$$

其中：χ_{1k} 表示第一条查询结果的第 i 个属性值，χ_{2k} 表示第二条查询结果的第 i 个属性值。

2）算法性能

为了保证网络时延的一致性，实验在相同网络环境下进行。通过计算算法提交和结果返回的时间间隔，来对其进行比较，从而比较模型性能。间隔时间越短，说明模型性能越好。由于DPMCA 过程可以独立运行，DPMCA 时间未计入实验所需的时间间隔。当查询规模 m 大于 20000 时，传统模型无法在有效时间内收敛，故分别设置查询规模 m 为 5000，8000 和 10000。

4.2　实验结果分析

1）数据无关性分析

通过关联性分析以及设置不同的最小支持度，可以得出如表 2 的结果。

表 2　数据无关性分析表

最小支持度	数据集包含的项	数据无关处理后的项
0.35	22	18
0.5	22	18
1.0	22	20
1.25	22	21

由表 2 可以看出，经过数据无关处理后，数据集的规模得到了有效的减少，会使之后的矩阵分解的计算量以及噪声的添加量减少。

2）数据可用性实验结果(欧氏距离)

表 3　数据可用性实验结果表

E	IQDPPBD	LRM	MM
0.5	92	107	110
0.75	95	92	95
1.0	65	79	83
1.25	66	65	70

由表 3 可以看出，由于采用自适应模型，低权限用户IQDPPBD 的查询结果精度与 ε 为 0.75 时几乎相当，较高权限用户与 ε 为 1.25 时相当。结果表明，IQDPPBD 模型可以实现较少噪声量的自适应添加以及较高查询精度的隐私保护。

3）算法性能实验结果(时间单位为 s)

表 4　算法性能实验结果表

m	IQDPPBD	LRM	MM
100	3	2	2
200	5	3	4
5000	97	117	121
8000	145	181	187
10000	183	251	260

由表 4 可以看出，当查询规模为 100，200，5000，8000，10000 时，模型的算法性能状况。其中，m 是查询规模，IQDPPBD 是本文提出的模型，LRM 和 MM 是传统模型。

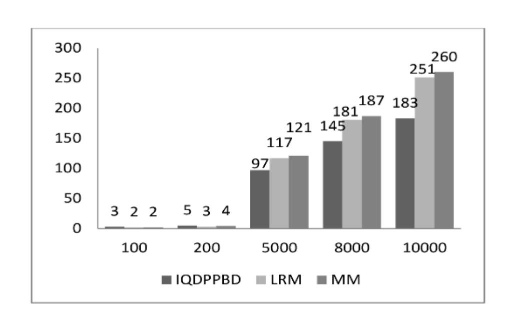


图 5　算法性能直方图

由图 5 可以看出，当查询规模分别为 5000，8000，10000时，本文提出的 IQDPPBD 模型的性能均优于 LRM[23]和 MM[22]模型。而当查询规模为 100，200 时，IQDPPBD 模型性能比 LRM和 MM 差。其中，图 5 中的横坐标表示查询规模 m，纵坐标表示运行时间（单位 s），用来衡量算法性能。

以上实验结果表明，IQDPPBD 模型可以自适应添加噪声。同时，当查询规模较大时，算法性能较已有模型有很大提高；当查询规模较小时，由于提出的模型对数据进行关联处理的时间相对查询时间而言较长，导致其性能没有传统模型好。因此IQDPPBD 模型可以满足大数据集中交互式线性查询隐私保护对系统响应和扩展性的严格要求。

5　结束语

在大规模数据集中，线性查询操作最基础且最频繁，交互式批量线性查询是线性查询的一种方式，由于其比普通的线性查询更难泄密且查询效率较高而被频繁使用，但已有的隐私保护模型很少研究对交互式批量线性查询的保护，基于此，本文提出了 IQDPPBD 模型。该模型针对交互式批量线性查询提出了隐私保护程度、数据可用性以及隐私保护模型效率的要求，通过挖掘记录属性的相关项，建立无关项构建的负载矩阵，用ADMM 模型对矩阵的求解进行优化，同时采用自适应模型进行隐私加噪，实现了差分隐私保护。实验采用 webdocs 数据集，运用分布式集群验证模型的并行性和可扩展性，从数据可用性和模型性能角度对比 IQDPPBD、LRM[23]和 MM[22]模型。结果表明，本文所提出的模型性能均优于传统隐私保护模型。另外，在不同隐私预算 ε 下，对隐私保护度以及数据准确率进行评估比较，证明了该模型的可行性。在未来的工作中，可以考虑将线性查询记录按内容特点进行分类，以便针对不同类型改进模型，在进一步保证数据准确性的同时优化矩阵机制的性能，从而使更大规模的数据集能快速收敛。该模型适用于对大规模数据集进行交互式线性查询隐私保护，当数据集较小时，算法性能反而低于传统模型。随着查询规模的增加，该模型的性能更优于传统模型。因此，在下一步的工作中将定量分析数据规模大小与该隐私保护模型下的交互式线性查询效率之间的关系。

参考文献：

[1] Dalenius T. Towards a methodology for statistical disclosure control [J].

Statistisk Tidskrift, 1977, 15 (2): 429-444.

[2] Chaudhuri K, Monteleoni C, Sarwate A D. Differentially private empirical risk minimization [J]. Journal of Machine Learning Research, 2011, 12 (2): 1069.

[3] Dwork C, Mcsherry F, Nissim K. Calibrating noise to sensitivity in private data analysis [J]. Proceedings of the VLDB Endowment, 2006, 7 (8): 637-648.

[4] Dwork C, Mcsherry F, Nissim K. Calibrating noise to sensitivity in private data analysis [C]// Proc of International Conference on Theory of Cryptography. [S. l.] : Springer-Verlag, 2006: 265-284.

[5] Proserpio D, Goldberg S, Mcsherry F. A platform for differentially private social graph analysis or, calibrating data to sensitivity in private data analysis [J]. arXiv: 1203. 3453 2013: 637-648.

[6] Cai Shulan, Zhang Zhongping, Song Li, *et al.* A k-anonymity model for protecting privacy [J]. Journal of Yanshan University, 2005.

[7] Abidi B, Yahia S B. Generating k-anonymous microdata by fuzzy possibilistic clustering [C]// Proc of International Conference on Database and Expert Systems Applications. [S. l.] : Springer, 2017: 3-17.

[8] Mezher A M, Álvarez A G, Rebollo-Monedero D, *et al.* Computational improvements in parallelized k-anonymous microaggregation of large databases [C]// Proc of IEEE, International Conference on Distributed Computing Systems. 2017, 258-264.

[9] Dargahi T, Ambrosin M, Conti M, *et al.* ABAKA: a novel attribute-based k-anonymous collaborative solution for LBSs [J]. Computer Communications, 2016, 85: 1-13.

[10] Casino F, Domingo-Ferrer J, Patsakis C, *et al.* A k-anonymous approach to privacy preserving collaborative filtering [M]. [S. l.] : Academic Press, 2015.

[11] Machanavajjhala A, Kifer D, Gehrke J. l-diversity: privacy beyound k-anonymity [C]// Proc of the 22nd International Conference on Data Engineering. 2006.

[12] Nie Weilin, Wang Cheng. Probability comprehension of differential privacy for privacy protection algorithms: A new measure [J]. International Journal of Wavelets Multiresolution & Information Processing, 2017.

[13] Wu Zhengang, Advance on privacy protection techniques for big data applications [J]. Telecommunications Network Technology, 2016 (2) .

[14] Wang Shuo, Sinnott R O. Protecting personal trajectories of social media users through differential privacy [J]. Computers & Security, 2017, 67: 142-163.

[15] Wang Jun, Liu Shubo, Li Yongkai. A review of differential privacy in individual data release [J]. International Journal of Distributed Sensor Networks, 2015, 2015 (9): 1-18.

[16] To H, Ghinita G, Fan Liyue, *et al.* Differentially private location protection for worker datasets in spatial crowdsourcing [J]. IEEE Trans on Mobile Computing, 2017, 16 (4): 934-949.

[17] Heinzedeml C, McWilliams B, Meinshausen N. Preserving differential privacy between features in distributed estimation [J/OL]. 2017. https://arxiv.org/pdf/1703. 00403. pdf.

[18] Shen Entong, Yu Ting. Mining frequent graph patterns with differential privacy [C]// Proc of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2013: 545-553.

[19] Wang Yu, Huang Zhenqi, Sayan M, *et al.* Differential privacy in linear distributed control systems: entropy minimizing mechanisms and performance trade-offs [J]. IEEE Trans on Control of Network Systems, 2017, PP (99): 1-1.

[20] Zhang Lin, Liu Yan, Wang Ruchuan, *et al.* Efficient privacy preserving classification construction model with differential privacy technology [J]. Journal of Systems Engineering and Electronics. 2017, 28 (1): 170-178.

[21] Li Chao, Hay M, Rastogi V, *et al.* Optimizing linear counting queries under differential privac [C]// Proc of ACM Symposium on Principles of Database Systems. 2010: 123-134.

[22] Li Chao, Miklau G, Hay M, *et al.* The matrix mechanism: optimizing linear counting queries under differential privacy [J]. International Journal on Very Large Data Bases, 2015, 24 (6): 757-781.

[23] Yuan Ganzhao, Zhang Zhenjie, Winslett M, *et al.* Low-rank mechanism [J]. Proceedings of the VLDB Endowment, 2012, 5 (11): 1352-1363.

[24] Hardt M, Talwar K, On the geometry of differential pri-vacy [C]// Proc of ACM Symposium on Theory of Computing. New York: ACM Press, 2010: 705-714.

[25] Bhaskara A, Dadush D, Krishnaswamy R, *et al.* Uncon-ditional differentially private mechanisms for linear queries [C]// Proc of Annual ACM Symposium on Theory of Computing. New York: ACM Press, 2012: 77–82.

[26] Yuan Jiawei, Tian Yifan. Practical privacy-preserving mapreduce based k-means clustering over large-scale dataset [J]. IEEE Trans on Cloud Computing, 2017, PP (99): 1.

[27] Jiang Huowen, Zhan Qinghua, Liu Wenjuan, *et al.* Clustering-anonymity approach for privacy preservation of graph data-publishing [J]. Journal of Software, 2017.

[28] Blum A, Ligett K, Roth A. A learning theory approach to noninteractive database privacy [J]. Proceeding of the ACM, 2011, 60 (2): 12-21.

[29] Hardt M, Rothblum G N. A multiplicative weights mec-hanism for privacy-preserving data analysis [C]// Proc of IEEE Symposium on Foundations of Computer Science. 2010: 61-70.

[30] Roth A, Roughgarden T. Interactive privacy via the median mechanism [C]// Proc of the 42nd ACM Symposium on Theory of Computing. New York: ACM Press, 2011: 765-774.

[31] Cao Hui, Liu Shubo, Zhao Renfang, *et al.* A privacy preserving model for energy internet base on differential privacy [C]// Proc of IEEE International Conference on Energy Internet. 2017: 204-209.